AFRL-IF-RS-TR-2006-306
Final Technical Report
October 2006

ON DEVELOPING THEORY AND APPLICATION OF COMMUNITY GENERATION

The Watson School of Engineering and Applied Science

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

STINFO FINAL REPORT

AIR FORCE RESEARCH LABORATORY
INFORMATION DIRECTORATE
ROME RESEARCH SITE
ROME, NEW YORK

NOTICE AND SIGNATURE PAGE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

This report was cleared for public release by the Air Force Research Laboratory Rome Research Site Public Affairs Office and is available to the general public, including foreign nationals. Copies may be obtained from the Defense Technical Information Center (DTIC) (http://www.dtic.mil).

AFRL-IF-RS-TR-2006-306 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

FOR THE DIRECTOR:

/s/	/s/
JOHN J. SALERNO .	JOSEPH CAMERA
Work Unit Manager	Chief, Information & Intelligence Exploitation Division Information Directorate

This report is published in the interest of scientific and technical information exchange, and its publication does not constitute the Government's approval or disapproval of its ideas or findings.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
OCT 06	Final	Jun 04 – Jun 06

4. TITLE AND SUBTITLE
ON DEVELOPING THEORY AND APPLICATION OF COMMUNITY GENERATION

5a. CONTRACT NUMBER
FA8750-04-1-0234

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
62702F

6. AUTHOR(S)
Zhongfei (Mark) Zhang

5d. PROJECT NUMBER
558B

5e. TASK NUMBER
IF

5f. WORK UNIT NUMBER
E1

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
The Watson School of Engineering and Applied Science
State University of New York (SUNY) Binghamton
P.O. Box 6000
Binghamton, New York 13902-6000

8. PERFORMING ORGANIZATION REPORT NUMBER
N/A

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Air Force Research Laboratory/IFEA
525 Brooks Rd
Rome NY 13441-4505

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSORING/MONITORING AGENCY REPORT NUMBER
AFRL-IF-RS-TR-2006-306

12. DISTRIBUTION AVAILABILITY STATEMENT
APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED. PA #06-718

13. SUPPLEMENTARY NOTES

14. ABSTRACT
The primary focus of this project was to develop a theory on community generation. Specific techniques as the Block Value Decomposition (BVD) and soft correspondence ensemble clustering frameworks, spectral relational clustering algorithm, and the general relation summary network model were designed and built. The BVD framework is a general framework for co-clustering dyadic relational data, which is a typical type of relational data in many applications. The soft correspondence ensemble clustering framework is a general framework for combining different clustering results together to deliver the optimal clustering result, which has many applications in distributed data mining and privacy-preserving data mining. The spectral relational clustering algorithm we have developed is a powerful relational data clustering algorithm that can be used for any type of relational data clustering. Finally, the relation summary network is the most general model that incorporates all the previous work as well as many existing models and algorithms in the literature which may be considered as the special cases of this model.

15. SUBJECT TERMS
Data Mining, Community Generation, Bi-Party Data Sets, Clustering, Block Value Decomposition

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			John J. Salerno
U	U	U	UL	24	19b. TELEPHONE NUMBER (*Include area code*)

Abstract

This is the final report of the project titled *On Developing Theory and Applications of Community Generation* supported by the Information Institute of US Air Force Research Laboratory under grant number FA8750-04-1-0234 technically managed by Dr. John Salerno at IFEA. The project begins in June, 2004, and ends in June, 2006 for the two years term with the total funding of $80,000 for the whole project. The project was awarded to SUNY Binghamton and was conducted at that campus. The project was extremely successful with the impressive achievements of four top conference publications (each with a very competitive acceptance rate) and several technology disclosures to the Tech Transfer office at SUNY Binghamton under review for possible further US patent applications. The specific technical achievements are documented in this report. Due to the success of this project with the excellent achievements, we have received substantial publicity in the relevant research community as well as in the related industrial and governmental sectors including Microsoft Research, IBM Research, Microsoft MSN, Google, Yahoo!, as well as DOE Berkeley/Lawrence Lab.

Table of Contents

List of Figures

List of Tables

Acknowledgements

Once again we would like to acknowledge the support in this grant from the Information Institute of AFRL. We thank Mr. John Graniero, Dr. Barry McKinney at the Information Institute for the advice and support. We thank the program manager, Dr. John Salerno, for his insightful advice and management.

1. Introduction

Data mining, ever since it was formally started as an independent research area, has been advancing rapidly over the last few years with extensive applications identified ranging from government to industries to even people's daily life [Han2006]. Community generation, as a recently emerging research topic in data mining area [Zhang2003, Salerno2004], has become one of the hottest research foci in the data mining community [Domingos2001, Mannila2002, Smyth2001, Long2005a, Long2006a, Long2006b]. The motivation for the research on community generation is due to the fact that the technologies developed from this research has found substantial application areas in governmental and industrial sectors, such as fraud detection [Goldberg1997, Jensen1997, Stolfo1997], crime investigation [Senator1995, Zhang2003, Salerno2004], sales promotion [Brin97a, Brin97b], social network analysis [Getoor2002, Kersting2000, Sarwar2001, Shardanand1995, Scott1991, Wassermann1994, Jensen2002, Glymour2001, Glymour1999], and Web mining [Aggarwal2001, White1996, Mack2002, Sarwar2001, Gibson1998], to just name a few. On the other hand, as of today, there is no well-established theory developed in the research on this topic. Given this context, this project aims to develop the theory on community generation as well as to identify the applications using the theory developed in this project.

Prior to the project, the PI has initiated the research on community generation in collaboration with his AFRL mentor, Dr. John Salerno, who is also the program manager of this project. In this preliminary research, we have identified the new paradigm of Uni-party Data Community Generation (UDCG), in contrast to the existing work in the literature on Bi-party Data Community Generation (BDCG), which is also referred to as relational data community generation. This preliminary work served as the foundation for this project, and received significant publication [Zhang2002, Zhang2003, Salerno2004].

This project is funded through Information Institute for the term between June, 2004, and June, 2006 for the total funding scale of $80,000. Under this project, one of the PI's PhD students, Bo Long, was supported. We have made excellent progress with very impressive achievements in this project [Long2005a, Long2005b, Long2006a, Long2006b]. Due to these impressive achievements, we have received substantial publicity in the data mining community and have received invitations for collaborations from the major industrial and governmental research and development organizations including Microsoft Research, IBM Research, Microsoft MSN, Google, Yahoo!, as well as DOE Berkeley/Lawrence Lab. Part of the technical achievements are being considered as the Tech Transfer Office of SUNY Binghamton for possible further applications to US Patents. In the following sections, we briefly summarize the major achievements in this project.

2. Block Value Decomposition for Dyadic Data Community Generation

The clustering is used in community generation in many disciplines and has a wide range of applications. In many applications, such as document clustering, collaborative filtering, and microarray analysis, the data can be formulated as a two-dimensional matrix representing a set of dyadic data. Dyadic data refer to a domain with two finite sets of objects in which observations are made for *dyads*, i.e., pairs with one element from either set. For the dyadic data in these applications, co-clustering both dimensions of the data matrix simultaneously is often more desirable than traditional one-way clustering. This is due to the fact that co-clustering takes the benefit of exploiting the duality between rows and columns to effectively deal with the high dimensional and sparse data that is typical in many applications. Moreover, there is an additional benefit for co-clustering to provide both row clusters and column clusters at same time. For example, we may be interested in simultaneously clustering genes and experimental conditions in bioinformatics applications, simultaneously clustering documents and words in text mining, simultaneously clustering users and movies in collaborative filtering.

In this work [Long2005a], we have developed a new co-clustering framework called Block Value Decomposition (BVD). The key idea is that the latent block structure in a two-dimensional dyadic data matrix $\boldsymbol{Z}$ can be explored by its triple decomposition. The dyadic data matrix is factorized into three components, the row-coefficient matrix $\boldsymbol{R}$, the block value matrix $\boldsymbol{B}$, and the column-coefficient matrix $\boldsymbol{C}$, as shown in Eq. 1. The coefficients denote the degrees of the rows and columns associated with their clusters and the block value matrix is an explicit and compact representation of the hidden block structure of the data matrix.

Under this framework, we develop a specific novel co-clustering algorithm for a special yet very popular case -- non-negative dyadic data that iteratively computes the three decomposition matrices based on the multiplicative updating rules derived from an objective criterion. By intertwining the row clusterings and the column clusterings at each iteration, the algorithm performs an implicitly adaptive dimensionality reduction, which works well for typical high-dimensional and sparse data in many data mining applications. The algorithm has been implemented in two cases, one for the asymmetric dyadic data matrix called NBVD, and the other for symmetric dyadic data matrix called symmetric NBVD. We have proven the correctness of the algorithm by showing that the algorithm is guaranteed to converge and have conducted extensive experimental evaluations to demonstrate the effectiveness and potential of the framework and the algorithms.

$$Z \approx RBC \tag{1}$$

Figures 1 and 2 shows two examples of the BVD applications and Figure 3 gives the conceptual illustration for the BVD operation.

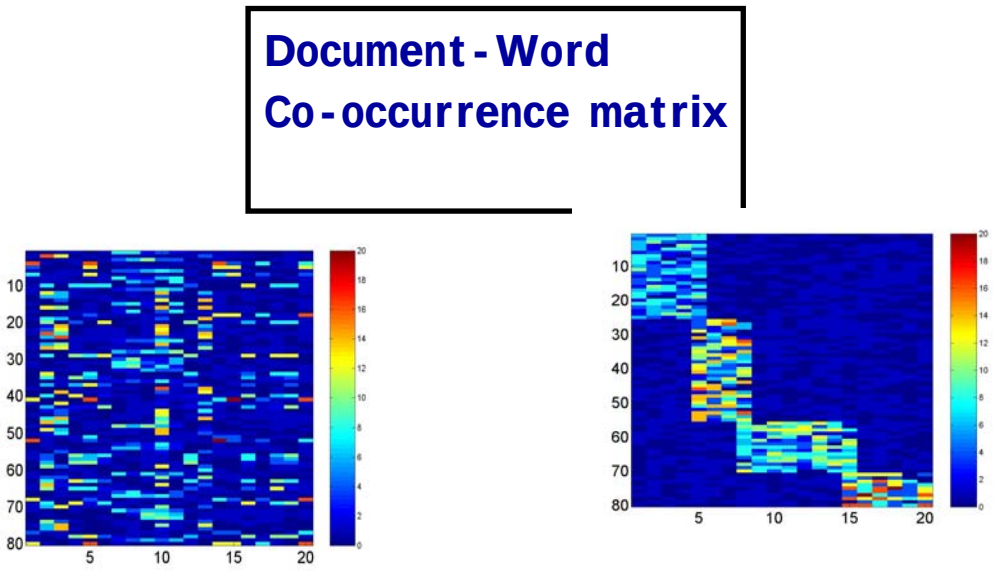

Figure 1 - BVD application example of document clustering

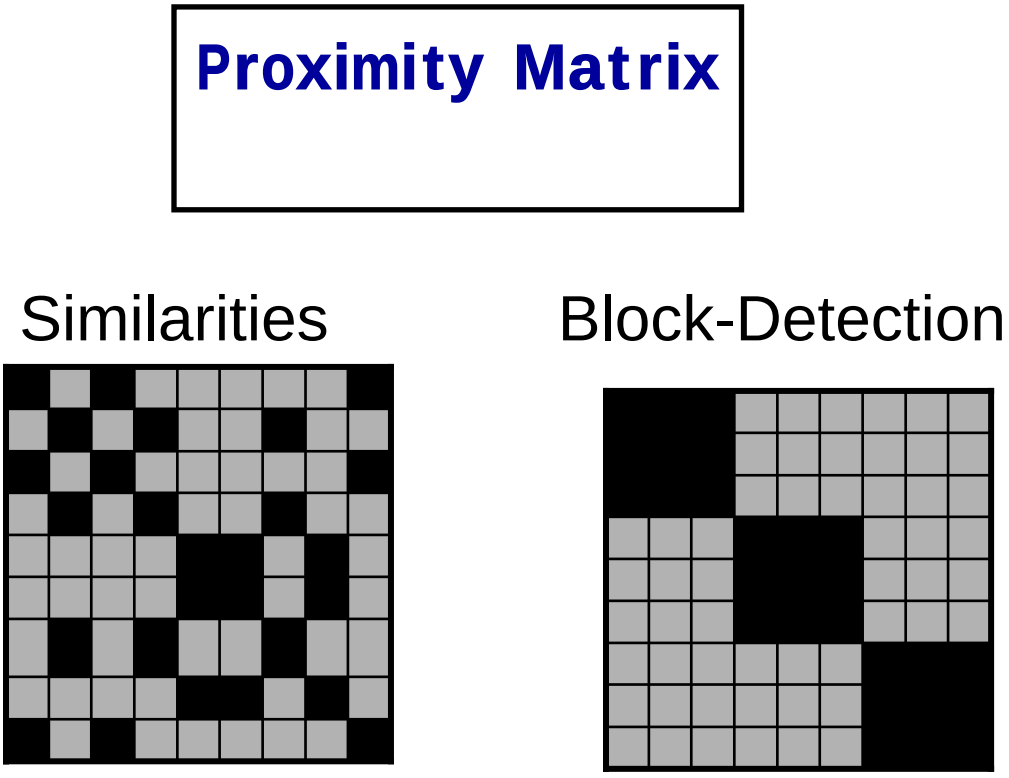

Figure 2 - Another example of BVD as clustering the proximity matrix

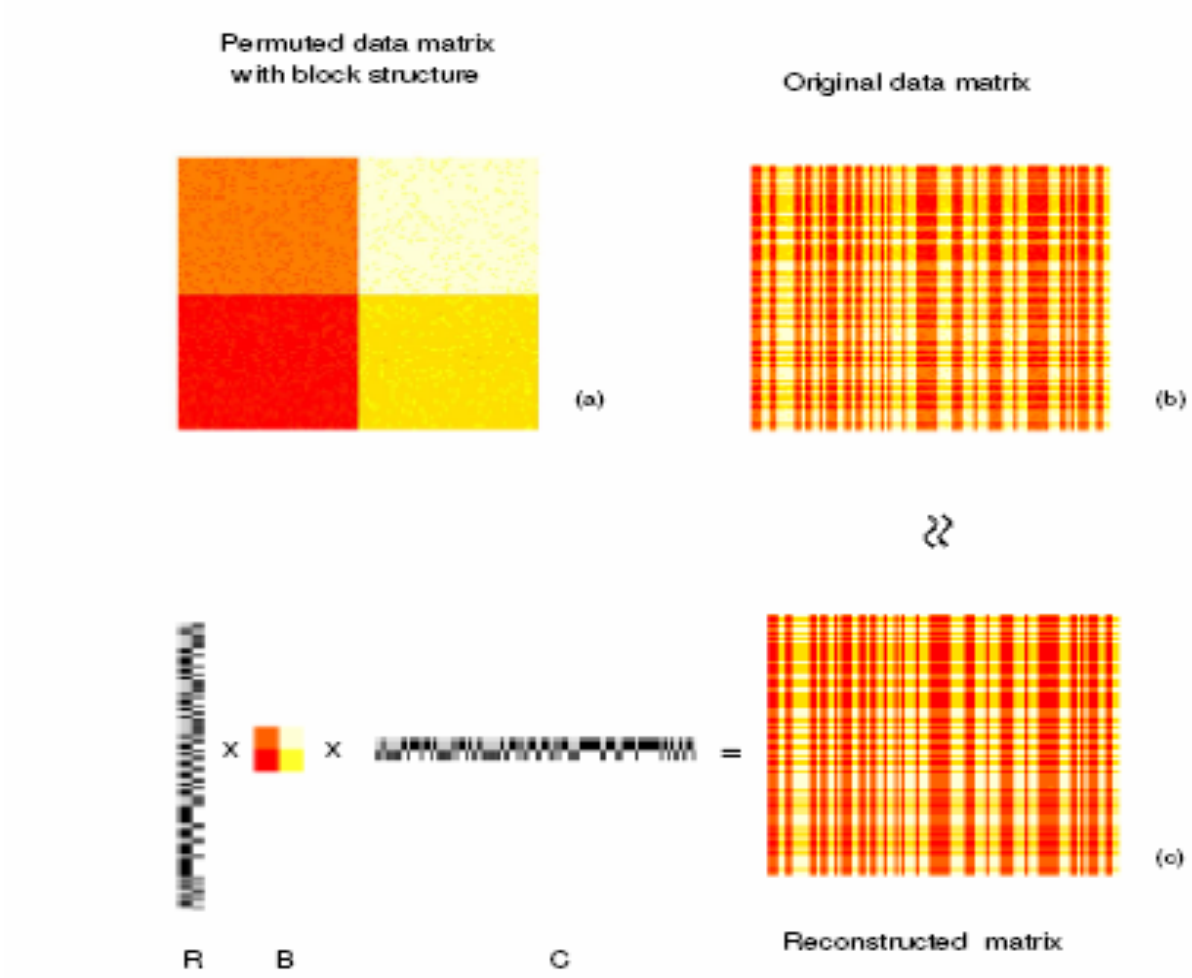

Figure 3 - Conceptual illustration of the BVD operation

We have used the 20 newsgroup data for evaluations of the BVD framework through the two cases of the algorithm. For evaluations of NBVD, we compare its performance with those of Non-negative Matrix Factorization (NMF) [Lee1999], Information-theoretic Co-Clustering (ICC) [Dhillon2003], and Iterative Double Clustering (IDC) [El-Yaniv2001]. Table 1 shows the performance comparison in terms of the precision values for three different 20 newsgroup data sets (binary, multi5, and multi10), which clearly demonstrates the superiority of BVD framework.

Table 1 - Performance comparison among NBVD and NMF, ICC, and IDC

	NBVD	NMF	ICC	IDC
Binary	0.95	0.91	0.96	0.85
Multi5	0.93	0.88	0.89	0.88
Multi10	0.67	0.60	0.54	0.55

For NBVD symmetric, we applied it to the proximity matrix partition problem and compared it with Normalized Cut (NC) [Shi2000] and Average Association (AA) [Zha2002]. Figure 4 documents the performance comparison which once again demonstrates the superiority of BVD framework.

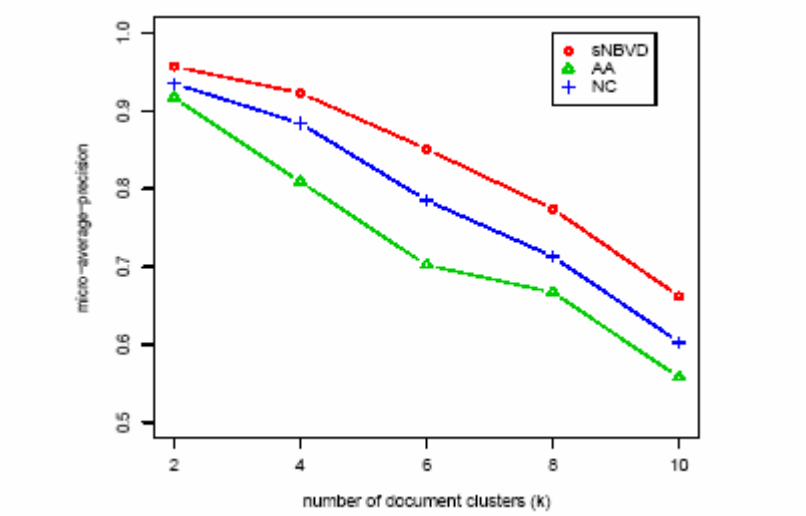

Figure 4 - Performance comparison among NBVD symmetric and NC and AA

3. Clustering Ensemble for Community Generation

After the community generation, there is a question of how good the community generation quality is. This is true when we apply different community generation algorithms to the same data collection, or even apply the same community generation algorithm with different parameters to the same data collection. Given such different community generation results for the same original data collection, which one shall we pick up? Without any a priori knowledge regarding the data collection, we really cannot decide which one is the best. The only solution we can take is to combine them all to obtain the best solution.

This approach becomes mandatory when we are in the scenario when the data collection is distributed over a network and each site of the network for community generation can only access to part of the whole global data collection that is distributed over the network. In this case, since each site is only visible to a part of the whole global data, it is expected that the community generation result at this site is not perfect. Therefore, given the individual community generation results at all the different sites, it is mandatory to combine all the results obtained at all the different sites together to secure the best result for the whole data collection.

Another common scenario similar to the distributed data collection is the privacy-preserving data mining in which each party is only visible to part of the whole data collection due to the privacy-preserving requirement and thus the community generation at this party can only be done based on the part of whole data collection. Consequently,

after all the parties obtain their own community generation results, it is necessary to obtain the community generation across all the parties for the best result. Figure 5 illustrates the scenario for clustering ensemble.

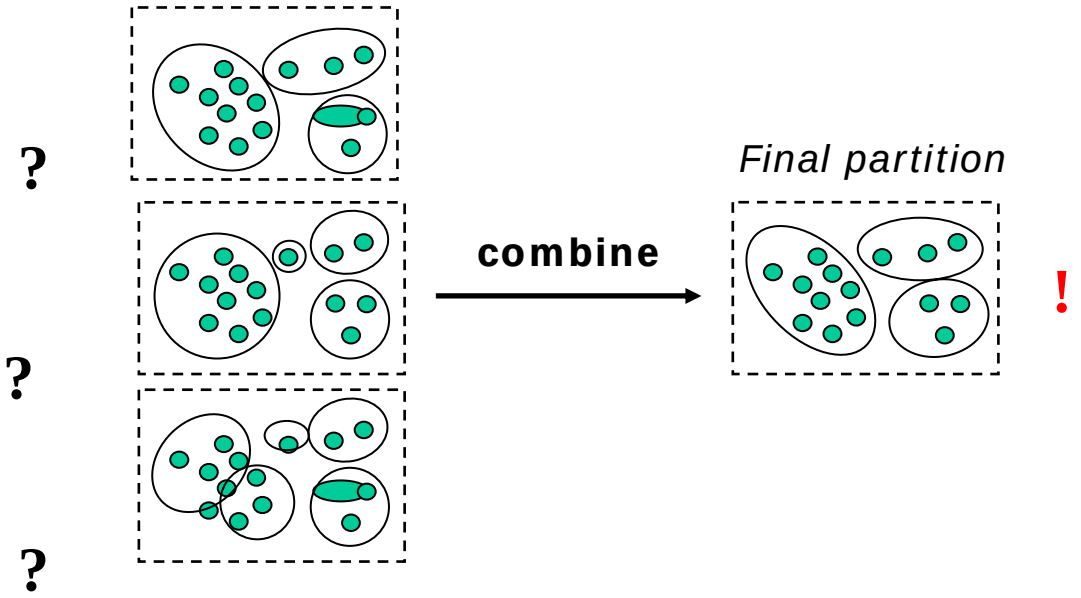

Figure 5 - Conceptual illustration of the community generation ensemble

Based on these motivations, we have developed a clustering ensemble approach called Soft Correspondence Ensemble Clustering (SCEC) for clustering based community generation ensemble [Long2005b]. Figure 6 gives the outlines of the SCEC algorithm.

Algorithm 1 SCEC($M^{(1)}, \ldots, M^{(k_r)}, k$)

1: Initialize $M, S^{(1)}, \ldots, S^{(r)}$.
2: **while** convergence criterion of M is not satisfied **do**
3: **for** $h = 1$ to r **do**
4: **while** convergence criterion of $S^{(h)}$ is not satisfied **do**
5: $S^{(h)} \leftarrow S^{(h)} \odot \frac{(M^{(h)})^T M + \beta k \mathbf{1}_{k_h k}}{D + \epsilon}$
6: **end while**
7: **end for**
8: $M = \frac{1}{r} \sum_{h=1}^{r} M^{(h)} S^{(h)}$
9: **end while**

Figure 6 - SCEC Algorithm outlines

The key technical challenge for community generation ensemble is to find the community correspondences between different individual community generation results. In the literature, all the existing methods have a strong assumption that individual communities generated in different results must satisfy one-to-one correspondence. This assumption is clearly hot true in many applications. The major significance and intellectual merit of SCEC is that in SCEC we propose soft correspondence instead of this kind of "hard" correspondence such that the one-to-one correspondence requirement is removed and thus the algorithm may be applied to any scenarios without such strong assumption. It can be shown that SCEC can always deliver the best ensemble solution.

Another advantage of SCEC is that due to its soft correspondence capability, it can handle the scenarios where there are missing attribute values in the data collection. SCEC can "automatically" take care of the missing values and is still able to deliver the best solution. Finally, SCEC promises to be an efficient algorithm in comparison with the existing methods in the literature.

We used the open source UCI data sets to evaluate the SCEC algorithm. Specifically, we used the IRIS, PENDIG, and ISOLET6 data sets. We compared SCEC with four existing methods from the literature: clustering based similarity partitioning algorithm (CSPA) [Strehl2002], metal clustering algorithm (MCLA) [Strehl2002], quadratic mutual information (QMI) [Topchy2003], and mixture model based ensemble clustering (MMEC) [Topchy2004], as well as the baseline k-means. For each comparison scenario for each data set, we try three different cases: random initialization (RI), random number (RN), and random subsets (RS). Figure 7 documents the overall performance comparison between SCEC and the existing methods and the k-means baseline for the nine scenarios (three data sets with the three initialization cases). From the figure, it is clear that SCEC outperforms all the existing methods and the baseline method in most of the cases.

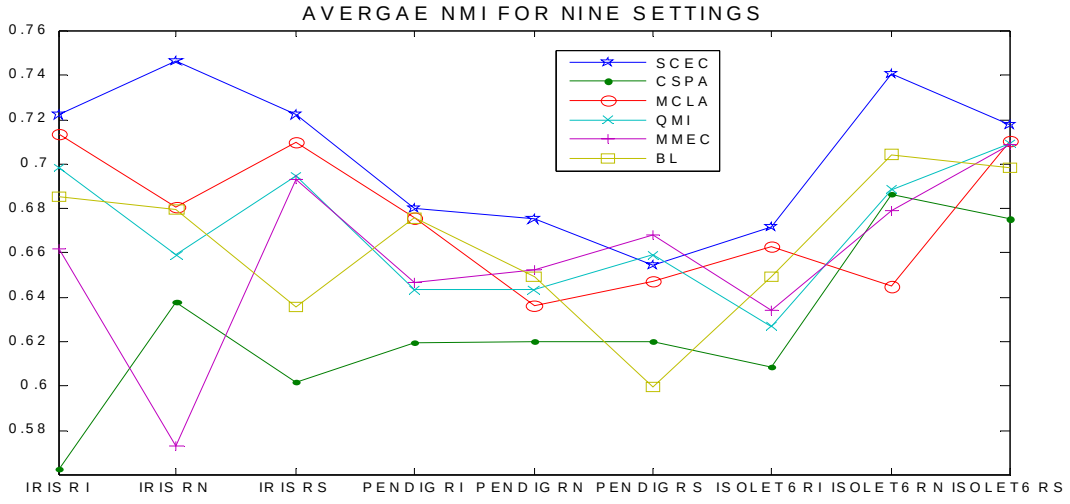

Figure 7 - Performance evaluations between SCEC and the existing methods and the baseline k-means under the nine different scenarios

4. Spectral Approach to Bi-party Data Community Generation

Most clustering based bi-party data community generation approaches in the literature focus on "flat" data in which each data object is represented as a fixed-length feature vector. However, many real-world data sets are much richer in structure, involving objects of multiple types that are related to each other, such as Web pages, search queries and Web users in a Web search system, and papers, key words, authors and conferences in a scientific publication domain. In such scenarios, using traditional methods to cluster each type of objects independently may not work well due to the following reasons.

First, to make use of relation information under the traditional clustering framework, the relation information needs to be transformed into features. In general, this transformation causes information loss and/or very high dimensional and sparse data. For example, if we represent the relations between Web pages and Web users as well as search queries as the features for the Web pages, this leads to a huge number of features with sparse values for each Web page. Second, traditional clustering approaches are unable to tackle with the interactions among the hidden structures of different types of objects, since they cluster data of single type based on static features. Note that the interactions could pass along the relations, i.e., there exists influence propagation in multi-type relational data. Third, in some machine learning applications, users are not only interested in the hidden structure for each type of objects, but also the global structure involving multi-types of objects. For example, in document clustering, except for document clusters and word clusters, the relationship between document clusters and word clusters is also useful information. It is difficult to discover such global structures by clustering each type of objects individually.

Therefore, multi-type relational data has presented a great challenge for traditional clustering approaches. In this study [Long2006a], first, we propose a general model, the collective factorization on related matrices, to discover the hidden structures of multi-types of objects based on both feature information and relation information. By clustering the multi-types of objects simultaneously, the model performs adaptive dimensionality reduction for each type of data. Through the related factorizations which share factors, the hidden structures of different types of objects could interact under the model. In addition to the cluster structures for each type of data, the model also provides information about the relation between clusters of different types of objects for identifying the global community structures in addition to the local clusterings. Figure 8 illustrates such an example in the application of Web mining.

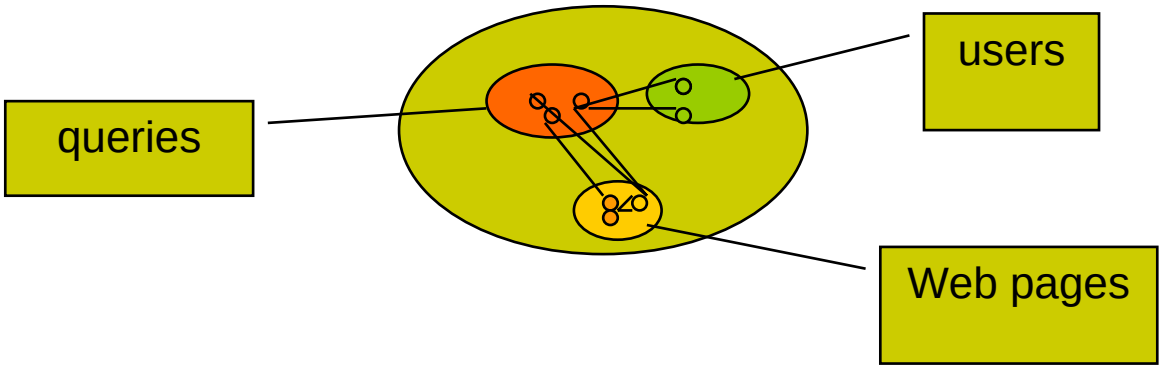

Figure 8 - An example of multi-type relational data mining for identifying the global community structures in addition to the local clusterings

Second, under this model, we derive a novel algorithm, the spectral relational clustering, to cluster multi-type interrelated data objects simultaneously. By iteratively embedding each type of data objects into low dimensional spaces, the algorithm benefits from the interactions among the hidden structures of different types of data objects. The algorithm has the simplicity of spectral clustering approaches but at the same time also applicable to relational data with various structures. Theoretic analysis and experimental results demonstrate the promise and effectiveness of the algorithm.

Third, we show that the existing spectral clustering algorithms can be considered as the special cases of the proposed model and algorithm. This provides a unified view to understand the connections among these algorithms.

Specifically, we propose the Collective Factorization on Related Matrices (CFRM) model. For each relation in the data collection of the multi-type relational data, we represent the relation as a related matrix $\boldsymbol{R}$. According to the BVD framework we have developed [Long2005a], this matrix can be decomposed into three components. Similarly, given a feature matrix $\boldsymbol{F}$, it can be considered as a special case of BVD and then can be decomposed into two components as the third one is an identity matrix. Consequently, we have Eq. 2 below for the general multi-type relational data CFRM model:

$$\min \sum_{1 \le i < j \le m} w_a^{(ij)} \| R^{(ij)} - C^{(i)} A^{(ij)} (C^{(j)})^T \|^2 + \sum_{1 \le i \le m} w_b^{(i)} \| F^{(i)} - C^{(i)} B^{(i)} \|^2 \qquad (2)$$

Based on this CFRM model, we have developed an algorithm called Spectral Relational Clustering (SRC) for clustering based community generation for the general multi-type relational data scenario. The technical significance and intellectual merits of SRC are that it is as simple as the traditional spectral clustering but at the same time can be applied to

relational data with various structures; in addition, it has the advantage of low dimension embedding during the clustering, and also it is efficient in comparison with the existing methods in the literature.

To extensively evaluate the performance of SRC, we use the 20 newsgroup data. We compare the performance of SRC with those of the existing methods in the literature: normalized cut (NC) [Shi2000], Bipartite Spectral Graph Partitioning (BSGP) [Dhillon2001], Mutual Reinforcement K-means (MRK) [Long2006a], and Consistent Bipartite Graph Co-partitioning (CBGC) [Gao2005]. Figure 9 documents the evaluations of the bi-partite graph scenario for the 20 newsgroup data set, and Figure 10 documents the evaluations of the tri-partite graph scenario for the same 20 newsgroup data set. It is noted that in some cases where the data sets are large, CBGC ran out of the memory and so the data were not available. From these figures, it is clear that SRC outperforms all the comparing methods in all the cases.

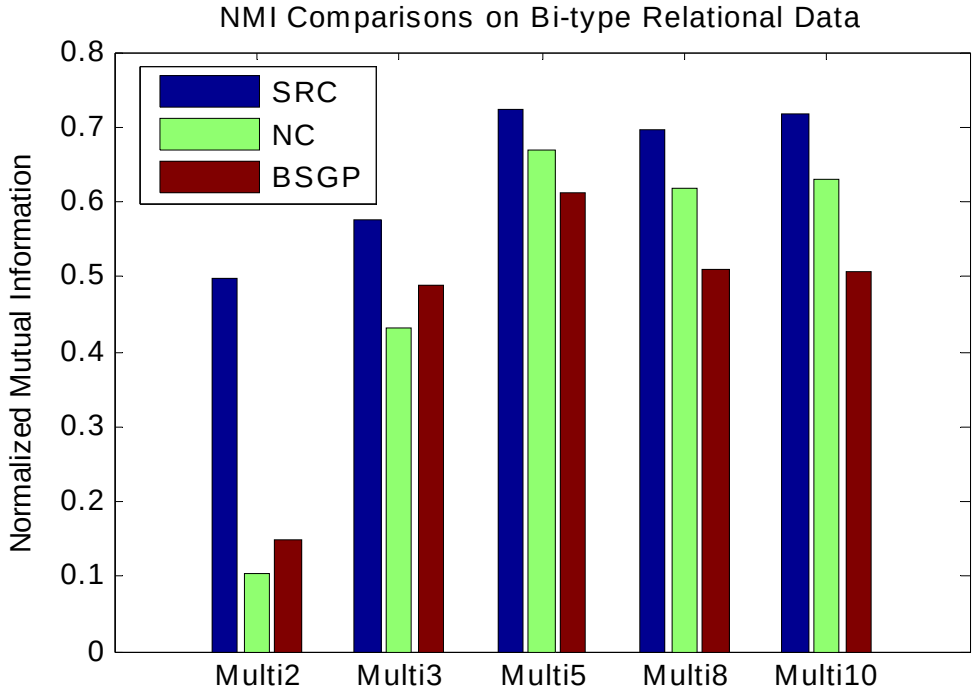

Figure 9 - Evaluations of SRC against NC and BSGP for the bi-partite graph scenario using the 20 newsgroup data set

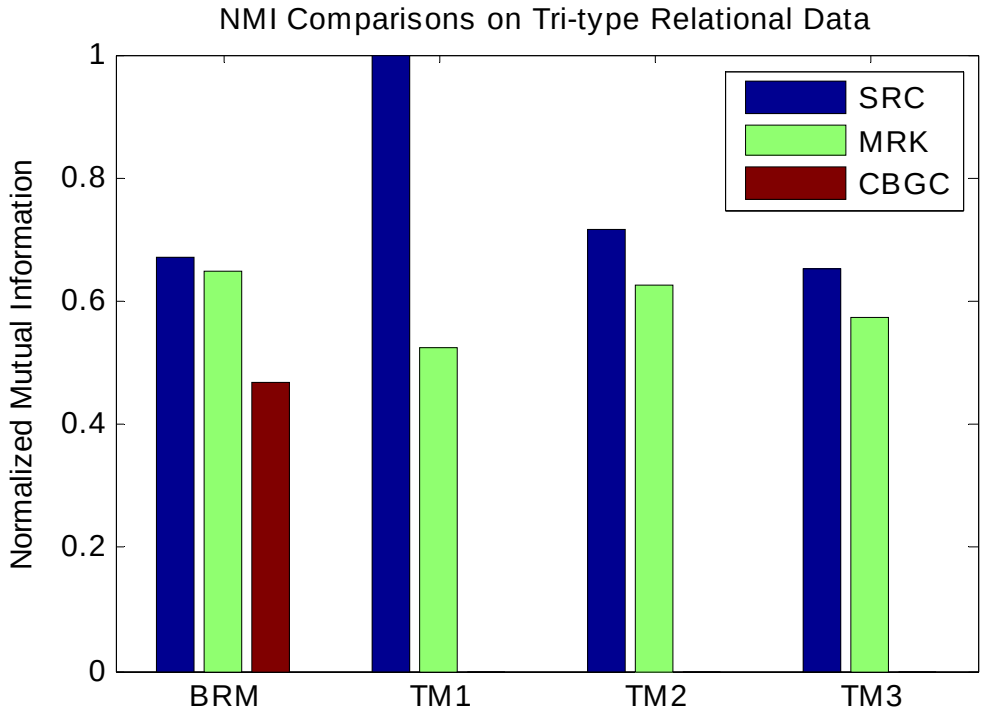

Figure 10 - Evaluations of SRC against MRK and CBGC for the tri-partite graph scenario using the 20 newsgroup data set

5. Relation Summary Network as a General Framework

Based on the above study [Long2006a], we went further to propose a more general framework. We first generalize the multi-type relational data as a k-partite graph.
An intuitive attempt to mine the hidden structures from k-partite graphs is applying existing graph partitioning approaches to k-partite graphs. This idea may work in some special and simple situations. However, in general, it is infeasible. First, the graph partitioning theory focuses on finding the best cuts of a graph under a certain criterion and it is very difficult to cut different type of relations (links) simultaneously to identify different hidden structures for different types of nodes. Second, by partitioning the whole k-partite graph into m subgraphs, one actually assumes that all different types of nodes have the same number of clusters m, which in general is not true. Third, by simply partitioning the whole graph into disjoint subgraphs, the resulting hidden structures are rough. For example, the clusters of different types of nodes are restricted to one-to-one associations.

Therefore, mining hidden structures from k-partite graphs has presented a great challenge to traditional unsupervised learning approaches. In this study [Long2006b], first we propose a general model, the relation summary network (RSN), to find the hidden structures (the local cluster structures and the global community structures) from a k-partite graph. The basic idea is to construct a new k-partite graph with hidden nodes, which "summarize" the link information in the original k-partite graph and make the hidden structures explicit, to approximate the original graph. The model provides a

principal framework for unsupervised learning on k-partite graphs of various structures. Second, under this model, based on the matrix representation of a k-partite graph we reformulate the graph approximation as an optimization problem of matrix approximation and derive an iterative algorithm to find the hidden structures from a k-partite graph under a broad range of distortion measures. By iteratively updating the cluster structures for each type of nodes, the algorithm takes advantage of the interactions among the cluster structures of different types of nodes and performs implicit adaptive feature reduction for each type of nodes. Experiments on both synthetic and real data sets demonstrate the promise and effectiveness of the proposed model and algorithm. Third, we also establish the connections between existing clustering approaches and the proposed model to provide a unified view to the clustering approaches.

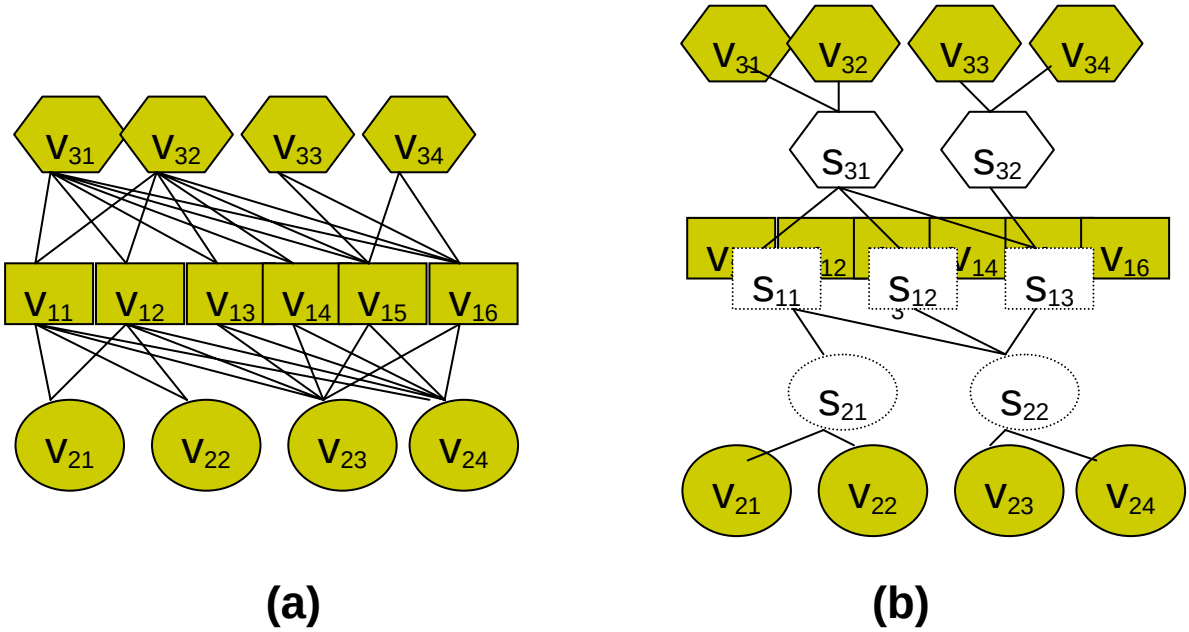

Figure 11 - An example of RSN model

Figure 11 illustrates an example of the RSN model where (a) is an original tri-partite graph and (b) is its corresponding RSN approximation. Based on this model, we use matrix representation for all the relations. Thus, according to the BVD framework [Long2005a], each relation can be decomposed into three components, and thus the key to the problem is to identify the closest approximation of the new graph with the summary nodes to the original graph.

This approach calls for using a distance metric to measure the "closeness" between the two graphs. We have developed a general algorithm under this RSN model that is applicable under a wide spectrum of distance functions w.r.t. different distance distributions. This algorithm is called Relation Summary Network with Bregman Divergence (RSN-BD). The algorithm is listed in Figure 12. Refer to [Long2006b] for the details of the algorithm.

Algorithm 1 Relation Summary Network with Bregman Divergences

Input: A k-partite graph $G = (V_1, \ldots, V_m, E)$, a Bregman divergence function D_ϕ, and m positive integers, $k_1, \ldots, k_m$.
Output: An RSN $G^s = (V_1, \ldots, V_m, S_1, \ldots, S_m, E^s)$.
Method:

1: Initialize G^s.
2: **repeat**
3: **for** $i = 1$ to m **do**
4: Update the edges between V_i and S_i according to Eq.(11).
5: **end for**
6: **for** each pair of $S_i \sim S_j$ where $1 \leq i < j \leq m$ **do**
7: Update the edges between S_i and S_j according to Eq.(13).
8: **end for**
9: **until** convergence

Figure 12 - RSN-BD algorithm

The advantages of RSN-BD includes: (1) Implicit adaptive dimensionality reduction through hidden nodes; (2) Applicable to K-partite graphs of various structures; (3) Applicable to graphs with different probabilistic distributions on their edges; and (4) efficient.

To evaluate the RSN model as well as the RSN-BD algorithm, we have implemented RSN-BD with four different Bregman divergence functions: Euclidean distance (RSN-ED), logistic loss (RSN-LL), generalized I-divergence (RSN-GI), and Itakura-Saito distance (RSN-IS). The first distance function corresponds to the normal distribution; the second corresponds to Bernoulli distribution; the third corresponds to Poisson distribution; and the fourth corresponds to the exponential distribution. We compare RSN with the K-means in the corresponding four cases (ED, LL, GI, and IS), BSGP [Dhillon2001], and CBGC [Gao2005].

We used simulated graph data as well as the real data for evaluations. The real data are the 20 newsgroup data set in which we generated the bi-partite and tri-partite graphs. In the simulated data, we generated Bernoulli, Poisson, and exponential distributions. Figure 13 documents the evaluations for the simulated bi-partite graphs under Bernoulli, Poisson, and exponential distributions among all the algorithms' performance. Figure 14 documents the performance comparisons among all the algorithms for three different bi-partite graphs for the real 20 newsgroup data set. Figure 15 documents the performance comparison among all the algorithms for the simulated exponential tri-partite graph and two other tri-partite graphs with the real 20 newsgroup data set. From these figures, it is clear that RSN model as well as the algorithm is superior to the comparing methods.

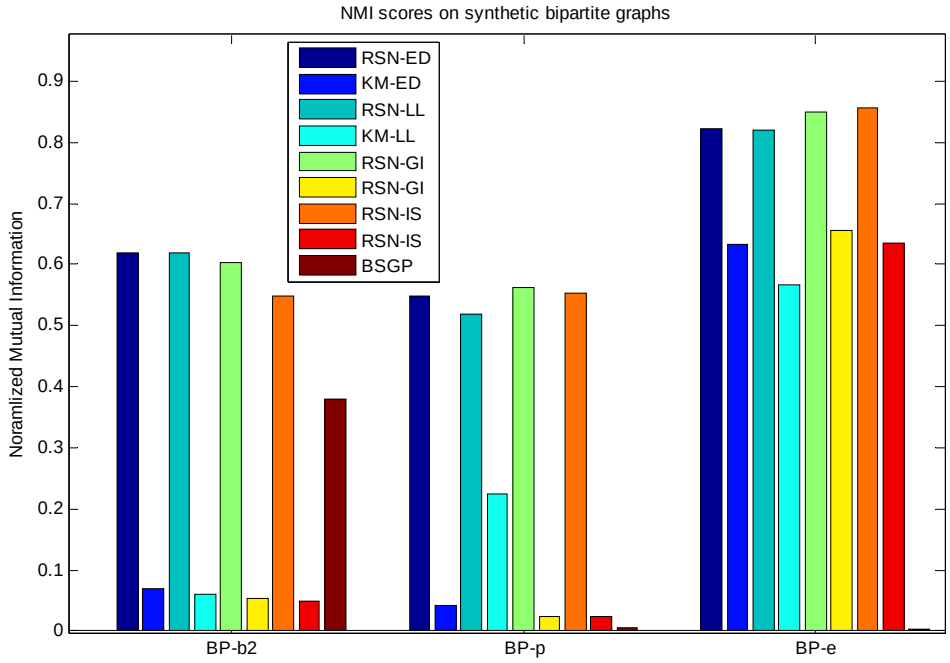

Figure 13 - Bi-partite graphs under Bernoulli, Poisson, and exponential distributions for the performance comparison among all the algorithms

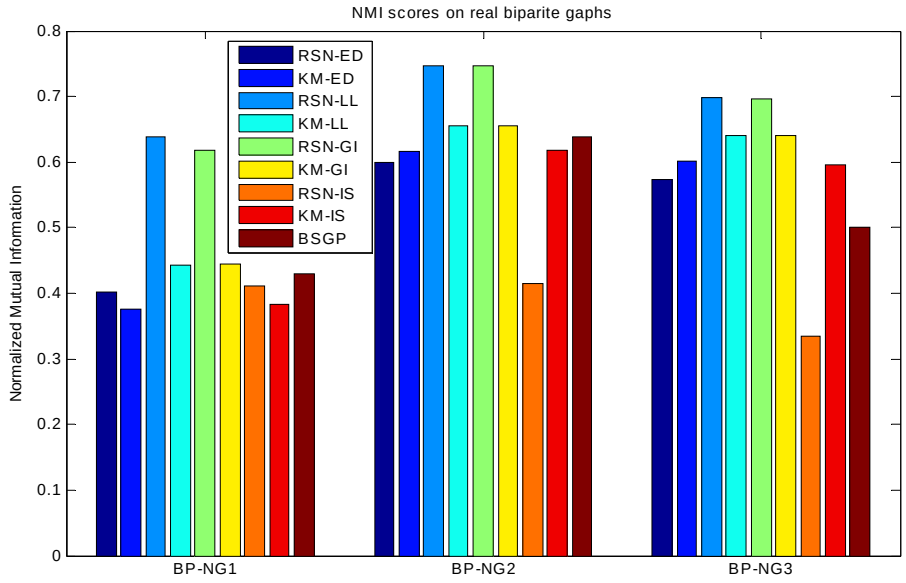

Figure 14 - Bi-partite graphs of the real 20 newsgroup data set for the performance comparison among all the algorithms

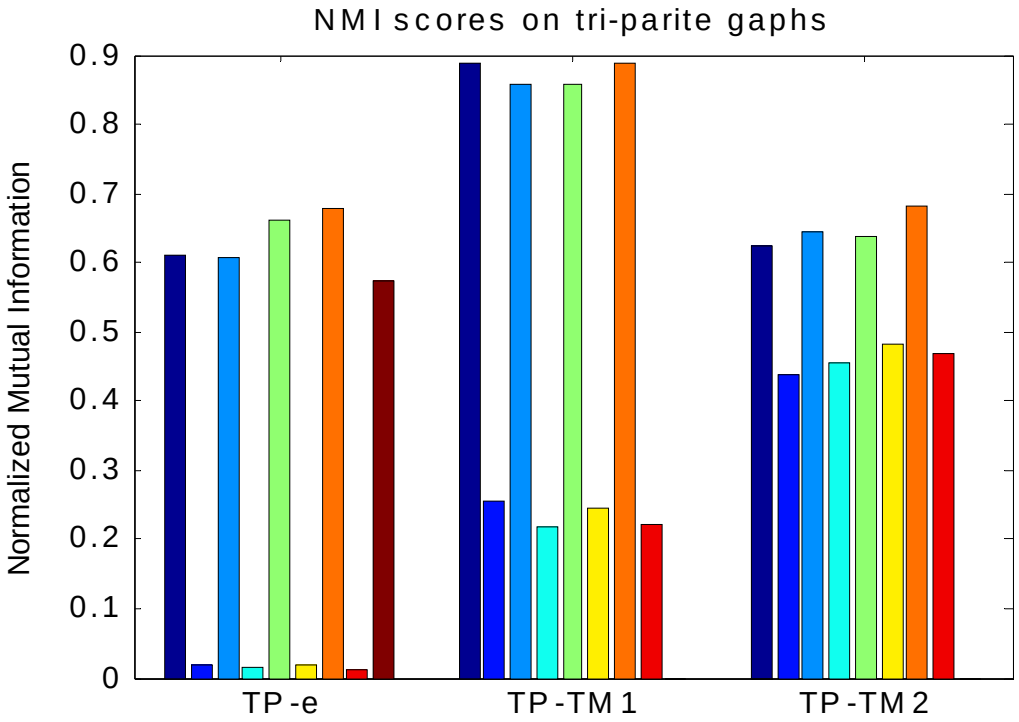

Figure 15 - Tri-partite graphs of simulated exponential distribution and two scenarios of the real 20 newsgroup data set for the performance comparison among all the algorithms

6. Conclusions

In this final report, we summarize the technical achievements made in the project, as well as the societal and broader impacts obtained through the publicity we have generated in this project. The project was extremely successful with substantial achievements and publicity generated. I hope that the techniques and the technologies generated from the research in this project shall be useful not only to the government, but also to the public. I hope that we will be continued to be supported in the future.

7. References

[Aggarwal2001] Aggarwal, C.C., F. Al-Garawi, and P.S. Yu, Intelligent crawling on the World Wide Web with arbitrary predicates, *Proc. ACM WWW*, 2001.

[Brin1997a] Brin, S., R. Motwani, and C. Silverstein, Beyond market basket: generalizing association rules to correlations, *Proc. ACM SIGMOD*, 1997.

[Brin1997b] Brin, S., R. Motwani, J.D. Ullman, and S. Tsur, Dynamic itemset counting and implication rules for market basket analysis, *Proc. ACM SIGMOD*, 1997.

[Dhillon2001] I.S. Dhillon, Co-clustering documents and words using bi-partite spectral graph partitioning, *KDD*, 2001.

[Dhillon2003] I.S. Dhillon, S. Mallela, and D.S. Modha, Information-theoretic co-clustering, *KDD*, 89 – 98, 2003.

[Domingos2001] Domingos, P. and D. Hulten, Catching up with the data: research issues in mining data streams, *Proc. Workshop on Research Issues in Data Mining and Knowledge Discovery*, 2001.

[El-Yaniv2001] R. El-Yaniv and O. Souroujon, Iterative double clustering for unsupervised and semi-supervised learning, *ECML*, 121 – 132, 2001.

[Gao2005] B. Gao, T.-Y. Liu, X. Zhang, Q.-S. Cheng, and W.-Y. Ma, Consistent biprtite graph co-partitioning for star-structured high-order heterogeneous data co-clustering, *KDD*, 2005.

[Getoor2002] Getoor, L., N. Friedman, D. Koller, and B. Taskar, Learning probabilistic models of link structure, *Journal of Machine Learning Research*, 2002.

[Gibson1998] Gibson, D., J. Kleinberg, and P. Raghavan, Inferring Web communities from link topoploy, *Proc. HyperText98*, 1998.

[Goldberg1997] Goldberg, H.G. and T.E. Senator, Break detection systems, *Proc. AAAI Workshop on AI Approaches to Fraud Detection and Risk Management*, AAAI Press, 1997.

[Glymour2001] Glymour, C. *The Mind's Arrows: Bayes Nets and Graphical Causal Models in Psychology*, MIT Press, 2001.

[Glymour1999] Glymour, C. and G. Cooper, *Causation, Computation and Discovery*, MIT/AAAI Press, 1999.

[Han2006] J. Han and M. Kamber, *Data Mining: concepts and techniques*, Morgan Kaufmann, 2nd Ed., 2006.

[Jensen1997] Jensen, D., Prospective assessment of AI technologies for fraud detection: A case study, *Proc. AAAI Workshop on AI Approaches to Fraud Detection and Risk Management*, 1997.

[Jensen2002] Jensen, D. and J. Neville, Data mining in social networks, *Symposium on Dynamic Social Network Modeling and Analysis, National Academy of Sciences*, National Academy Press, 2002.

[Kersting2000] Kersting, K. and L.D. Raedt, Bayesian logic programs, *Proc. 10th International Conf. Inductive Logic Programming*, 138 – 155, 2000.

[Lee1999] D.D. Lee and H.S. Seung, Learning the parts of objects by non-negative matrix factorization, *Nature*, 401:788 – 791, 1999

[Long2005a] B. Long, Z. Zhang, and P.S. Yu, Co-clustering by block value decomposition, *KDD*, 2005

[Long2005b] B. Long, Z. Zhang, and P.S. Yu, Combining multiple clusterings by soft correspondence, *ICDM*, 2005.

[Long2006a] B. Long, Z. Zhang, X. Wu, and P.S. Yu, Spectral clustering for multi-type relational data, *ICML*, 2006

[Long2006b] B. Long, X. Wu, Z. Zhang, and P.S. Yu, Unsupervised learning on k-partite graphs, *KDD*, 2006

[Mack2002] Mack, R. and M. Hehenberger, Text-based knowledge discovery: search and mining of life-sciences documents, *Drug Discovery Today*, 7(11), Suppl.: S89-S98, 2002.
[Mannila2002] Mannila, H., Local and global methods in data mining: basic techniques and open problems, *Proc. 29th International Colloquium on Automata, Languages, and Programming*, 2002.

[Salerno2004] J. Salerno, Z. Zhang, R. Lewin, and M. Decker, Discovering social groups without having relational data, *Proc. SPIE International Conf. on Data Mining and Knowledge Discovery: Theory, Tools, and Technology*, SPIE Press, 5433: 33 – 40, 2004.

[Sarwar2001] Sarwar, B., G. Karypis, J. Konstan, and J. Riedl, Item-based collaborative filtering recommendation algorithms, *Proc. ACM WWW*, 2001.

[Scott1991] Scott, J. *Social Network Analysis: A handbook*, SAGE Publications, 1991.
[Senator1995] Senator, T., H. Goldberg, J. Wooton, A. Cottini, A. Umar, C. Klinger, W. Llamas, M. Marrone, and R. Wong, The FinCEN artificial intelligence system: identifying potential money laundering from reports of large cash transactions, *Proc. The 7th Conf. Innovative Applications of AI*, August, 1995.

[Shardanand1995] Shardanand, U. and P. Maes, Social information filtering: algorithms for automating "world of mouth", *Proc. ACM CHI*, 1995.

[Shi2000] J. Shi and J. Malik, Normalized cuts and image segmentation, *IEEE T-PAMI*, 22(8):888-905, 2000.
[Smyth2001] Smyth, P., Breaking out of the black-box: research challenges in data mining, *Proc. Workshop on Research Issues in Data Mining and Knowledge Discovery*, 2001.

[Stolfo1997] Stolfo, S., D. Fan, A. Prodromidis, W. Lee, S. Tselepsis, and P. Chan, Agent-based Fraud and Intrusion Detection in Financial Information Systems, http://www.cs.columbia.edu/ sal/JAM/PROJECT, 1997.

[Strehl2002] A. Strehl and J. Ghosh, Cluster ensemble – a knowledge reuse framework for combining partitionings, *AAAI*, 2002.

[Topchy2003] A. Topchy, A.K. Jain, and W. Punch, Combining multiple weak clusterings, *ICDM*, 2003

[Topchy2004] A. Topchy, A.K. Jain, and W. Punch, A mixture model for clustering ensembles, *AIAM Data Mining*, 2004.

[Wassermann1994] Wassermann, S. and K. Faust, *Social Network Analysis: Methods and Applications*, Cambridge University Press, 1994.

[White1996] White, D.A. and R. Jain, Similarity indexing with the SS-tree, *Proc. IEEE International Conference on Data Engineering*, 1996.

[Zha2002] H. Zha, C. Ding, M. Gu, X. He, and H. Simon, Spectral relaxation for k-means clustering, *NIPS*, 14, 2002.

[Zhang2002] Zhang, Z. J.J. Salerno, M. Regan, and D. Cutler, Using data mining techniques for building fusion models, *Proc. SPIE International Conf. on Data Mining and Knowledge Discovery: Theory, Tools, and Technology V*, SPIE Press, 5098: 174 – 184, 2002.

[Zhang2003] Zhang, Z., J.J. Salerno, P.S. Yu, J. Hua, R. Zhang, M. Regan, and D. Cutler, Applying data mining in investigating money laundering crimes, *Proc. ACM KDD*, 2003.